A study of the effectiveness of articulatory strokes for phonemic recognition

Carlos Molina[1], Sungbok Lee[2], Shrikanth Narayanan[2], and Néstor Becerra Yoma[1]

[1] Speech Processing and Transmission Laboratory

Department of Electrical Engineering, Universidad de Chile, Santiago, Chile

[2] Signal Analysis and Interpretation Laboratory

Viterbi School of Engineering, University of Southern California, USA

Abstract

This paper explores a framework to incorporate articulatory movement information into a classical ASR scheme based on the concept of articulatory stroke. Articulatory stroke is a geometrical segmental unit which corresponds to a target approaching-releasing articulatory gesture. It has been shown that critical and non-critical (i.e., secondary or dummy) articulatory gestures can be classified with about 88% accuracy using the stroke parameters. Phonetic recognition accuracy is also investigated by augmenting the conventional MFCC features with the articulatory stroke features (obtained using the MOCHA corpus). It is found that the phonetic recognition accuracy increases 15% with respect to the best result using the ordinary MFCC parameters only. This provides supporting evidence for the usefulness of the articulatory stroke representation of articulatory movements not only for speech production description but also for automatic speech recognition.

Index Terms: articulatory stroke, multiple features integration; phoneme recognition; speech production

1. Introduction

One of the main problems faced by automatic speech recognition (ASR) technology in real applications is the mismatch between training and testing conditions. This problem has been tackled with model adaptation/compensation techniques or with methods for noise removal from the corrupted signal. Conventional adaptation/compensation techniques (e.g. MAP and MLLR) dramatically degrade when only a few adapting utterances are available [1][2][3]. One problem is the number of parameters that need to be re-estimated: the higher the number of model parameters, the larger the amount of required adaptation data. On the other hand, noise removal approaches usually require an accurate estimation of additive noise or channel distortion. It is worth highlighting that despite of all the techniques proposed so far, the problem of mismatch robustness in speech recognition is still open.

Articulatory domain analysis is a promising venue to improve the performance of ASR especially in mismatched scenarios (e.g., environment, channel noise, etc.)[4][5][6]. For example, in [4] a parallel method to extract information that combines conventional acoustic and articulatory features (AFs) shows promising results in noisy environment at low SNRs. In [5], a scheme of modeling separately the manner and place of articulation for sub-word units in speech recognition reports improvement of 10% in WER. Also in [6], a system based on a combination of the widely popular mel frequency cepstral coefficients (MFCCs) and AFs to improve phoneme recognition accuracy led to significant reductions in WER.

This paper makes use of a particular segmental analysis of articulatory movements previously published in [7] where "articulatory stroke", AS, units are defined. AS aims to provide a reliable representation of the speech articulatory movements. It is based on the maximum and minimum local points of curvature. In this paper, the AS is defined as:

- An articulatory motion segmented by two succeeding points of local minimum curvature

This definition of AS is employed to segment the measured articulatory trajectories. AS can be classified as critical or dummy (non critical) depending on which phoneme produces the movement. According to the gestural dictionary in [8] it is possible to know which articulation is critical given a pronounced phoneme. For example, to produce the sound corresponding to the fricative /s/ the tongue tip movement is critical. On the other hand, when the articulatory movement is not critical given a pronounced phoneme it is classified as dummy.

In this paper an automatic classification of AS and an automatic AS feature based phoneme recognition are proposed. The classification scheme assumes that the recognized alignment gives the phoneme at each time. It is important to label each AS as critical or dummy in the training database. GMMs are used to model the two types of AS and the acoustic phoneme representation based on AS features. The maximum likelihood criterion gives the decision to classify each AS and to recognize the phonemes. The scheme presented here establishes a straightforward mechanism to map the acoustic phonetic representation with an articulatory segment characterized with AS features. The method is validated based on the accuracy rate on an automatic phoneme recognition task.

2. Articulatory strokes classification

2.1. Component of Articulatory Strokes

Articulatory strokes can be decomposed into three main components: approach; turning point; and, release (Fig.1). The "approach" component corresponds to the previous movement to reach a target or "turning point" which is when the articulatory trajectory changes the current direction. "Release" feature is related to the movement necessary to finalize the articulation. The first step adopted here attempts to associate each AS with an acoustic frame. It is worth mentioning that the acoustic-phonetic segmentation is provided by the database employed here (MOCHA, see Section 4).

As described in [7], AS units are allocated in the time domain when the turning positions take place in the articulatory movement trajectory. In other words, turning position at frame t will be mapped to the corresponding phoneme. If more than one stroke is aligned with the same phoneme, the most similar turning position to the mean turning one associated to the phoneme is selected. This paper presents results on: evaluation of AS recognition error with consonants where tongue tip sensor is critical (Table 1); and, evaluation of AS-based phoneme recognition accuracy. The main hypothesis considered here is the fact that the information provided by an AS when it is critical is more representative and better correlated to the articulatory gesture than when it is non-critical (dummy).

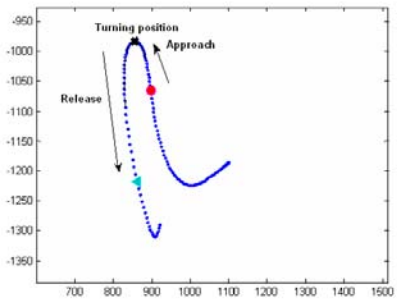

Figure 1: *Components of an AS: turning position; release; and, approach segment.*

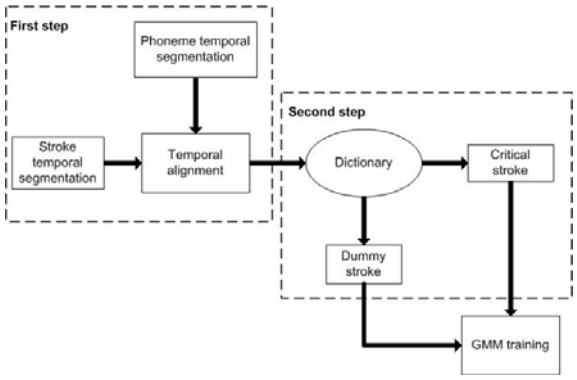

Figure 2: *Block diagram of the training scheme to estimate the critical and dummy GMMs*

2.2. Automatic strokes classification

In order to characterize ASs the same features employed in [7] are used here. The features vector is composed of: the mean of the velocity and acceleration into the stroke; the turning position; the maximum value of the curvature into the stroke; and, the duration of the stroke. Then, defining the AS feature vector as $\vec{V}_{sf}$, and, $\Phi_c(\cdot)$ and $\Phi_d(\cdot)$ as the pdf's of the critical and dummy strokes, respectively, the automatic stroke classification can be obtained using Bayes rule by solving:

$$\tilde{S} = \arg\max_{S}\left\{\Pr\left(\Phi_s\left|\vec{V}_{sf}\right.\right)\right\} = \arg\max_{S}\left\{\Pr\left(\vec{V}_{sf}\left|\Phi_s\right.\right)\right\} \quad (1)$$

where $S = \begin{cases} d\text{:dummy stroke} \\ c\text{:critical stroke} \end{cases}$

Figure 2 shows the scheme to train the GMMs associated with critical and dummy strokes. First, as mentioned in Sec. 2.1, the ASs are aligned to turning positions. Then, each AS is labeled as critical or dummy by employing the gestural dictionary. Finally, the extracted AS features are used to train the GMMs. In Fig. 3 the procedure to classify strokes based on the MAP criterion is described.

3. Articulatory based phoneme recognition

It is important to point out that the articulatory phoneme recognition is tested only with critical strokes. As mentioned before, the information provided by critical ASs is more representative and better correlated to the acoustic signal than those that are classified as dummy. The AS-based automatic phoneme recognition can be carried out in a supervised or unsupervised way. In supervised training the critical and dummy classification are done by using the gestural dictionary [8]. In contrast, unsupervised training makes use of the automatic classification presented in the previous section. As mentioned above, the experiments presented here take into consideration only the consonants which are labeled as critical for tongue tip sensor.

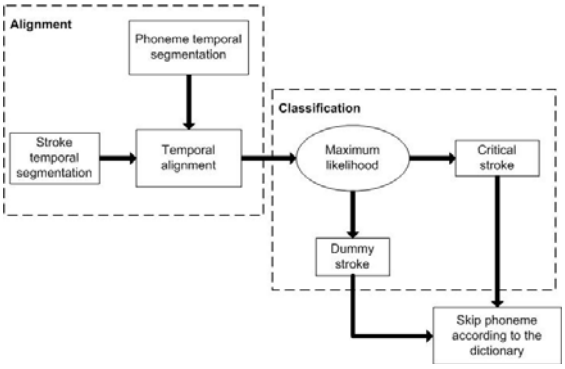

Figure 3: *Block diagram of the testing scheme to classify an AS.*

In this task, AS feature vectors are composed of the three sensors localized in the tongue: "tongue tip"; "tongue dorsum"; and, "tongue upper". Thus, if $\vec{V}_{sf}$ represents the feature vectors, then $\vec{V}_{sf} = \left\{\left(X_{TT},Y_{TT}\right),\left(X_{TB},Y_{TB}\right),\left(X_{TD},Y_{TD}\right)\right\}$ where $\left(X_{TT},Y_{TT}\right)$, $\left(X_{TB},Y_{TB}\right)$, and $\left(X_{TD},Y_{TD}\right)$ indicate the turning position of tongue tip, tongue body, and tongue dorsum, respectively. The recognized phoneme is obtained by means of the maximum likelihood criterion and then applying the Bayes rule:

$$\tilde{\lambda} = \arg\max_{\lambda}\left\{\Pr\left(\Phi_\lambda\left|\vec{V}_{sf}\right.\right)\right\} = \arg\max_{\lambda}\left\{\Pr\left(\vec{V}_{sf}\left|\Phi_\lambda\right.\right)\right\} \quad (2)$$

where λ can be any consonant labeled as critical according to the tongue tip sensor defined in the gestural dictionary; and, Φ_λ is the GMM of the phoneme λ.

Table 1: *Consonants where tongue tip sensor is critical according to [8].*

Consonants			
'ch'	'jh'	's'	'z'
'd'	'l'	'sh'	'zh'
'dh'	'n'	't'	
'hh'	'r'	'th'	

3.1. Combining Articulatory features and MFCCs

The first method explored here makes use of a feature vector that includes both MFCC and AS features. Thus, the feature vector $\vec{V}_{sf}$ in (2) is replaced with $\vec{V}$, where $\vec{V}$ is defined as $\vec{V} = \left\{\left(X_{TT},Y_{TT}\right),\left(X_{TB},Y_{TB}\right),\left(X_{TD},Y_{TD}\right),MFCC_1,...,MFCC_N\right\}$ and N corresponds to the number of MFCC. The recognized phoneme is obtained as shown in (2). On the other hand, following multi classification system (MCS) theory [9][9], it is possible to combine AS features and MFCC as a function of individual pdf's. In fact, if $\Pr\left(\Phi_\lambda\left|\vec{V}\right.\right)$ is equal to $\Pr\left(\Phi_\lambda\left|\vec{V}_{sf},\vec{V}_{MFCC}\right.\right)$, where $\vec{V}_{MFCC}$ represents the feature vector based on MFCC, then the joint probability $\Pr\left(\Phi_\lambda\left|\vec{V}_{sf},\vec{V}_{MFCC}\right.\right)$ could be approximated by employing the mean rule, shown in (3), or product rule approximations [9].

$$\Pr\left(\Phi_\lambda\left|\vec{V}_{sf},\vec{V}_{MFCC}\right.\right) = \Pr\left(\Phi_\lambda\left|\vec{V}_{sf}\right.\right) + \Pr\left(\Phi_\lambda\left|\vec{V}_{MFCC}\right.\right) \quad (3)$$

Then, assuming the approximation in (3), the optimal phoneme according to (2) is given by:

$$\tilde{\lambda} = \arg\max_{\lambda}\left\{\Pr\left(\Phi_{\lambda}\left|\vec{V}\right.\right)\right\} \approx \arg\max_{\lambda}\left\{\Pr\left(\Phi_{\lambda}\left|\vec{V}_{sf}\right.\right) + \Pr\left(\Phi_{\lambda}\left|\vec{V}_{MFCC}\right.\right)\right\} \quad (4)$$

3.2. Database description

The experiments presented here are based on the MOCHA database [10]. MOCHA is an EMA (Electromagnetic articulography) database of read speech by two speakers (female and male) of British English where each speaker produced 461 utterances. The acoustic features are 12 MFCCs computed every 10-millisecond interval. A 20 millisecond Hamming window is employed. As mentioned above, critical and dummy ASs are modeled with GMM and tested with variable numbers of Gaussians, N_{mix}. The Maximum likelihood criterion is employed to classify each AS.

3.3. Experiments

The experiments reported here attempt to evaluate the automatic stroke classification accuracy. In this sense, the data subset BT-Train is defined as the 361 utterances produced by one of the two speakers that are employed to train the GMM in (1). Two testing sets are employed:

- BT-1: corresponds to 100 utterances produced by the same training speaker.
- BT-2: corresponds to 100 utterances produced by the second speaker.

Both, BT-1 and BT-2 utterances employ the same transcription. The accuracy rate of the phoneme recognition task using articulatory features was evaluated employing either automatic stroke classification or by means of the dictionary presented in [8]. Finally, experiments that take into consideration MCS criteria in order to combine conventional MFCC parameters with the articulatory features described here are presented.

4. Results and Discussion

Tables 2-3 present the accuracy rate for AS detection (%) with different number of Gaussian components. The results reported here show that the best accuracy rate is observed when N_{mix} is equal to 8 and 4 for BT-1 and BT-2, respectively. Also, the automatic stroke classification accuracy rate varies widely depending on the test data set (88.36% and 64.72% for BT-1 and BT-2).

Table 2: *Accuracy rate (%) in the critical and dummy detection employing the AFs described in section 2 and testing with BT-1.*

Gaussian mixture number	Accuracy (%)
2	88.33
4	88.05
8	88.36
16	86.79

Table 3: *Accuracy rate (%) in the critical and dummy detection employing the AFs described in section 2 and testing with BT-2.*

Gaussian mixture number	Accuracy (%)
2	63.37
4	64.72
8	61.10
16	60.20

The reduction in classification accuracy by 29% on average when BT-1 is replaced with BT-2 reflects the fact that the articulatory features depend on the speaker's physiological characteristics. In order to avoid the speaker mismatch, the automatic phoneme recognition task only considers experiments with the same training and testing speaker. It is worth mentioning that speaker-independent representation can be devised based on a fixed landmark point such as the jaw and jaw-based coordinate system.

The baseline system corresponds to the automatic phoneme recognition (APR) by using the conventional MFCC parameters as feature vector. The training and testing dataset correspond to BT-Train and BT-1, respectively. The recognition accuracy versus N_{mix} with the baseline system is showed in Table 4. The MFCC based APR can lead to an accuracy rate of 53.1% with $N_{mix} = 4$.

Table 5 presents results by using the AFs described above and the unsupervised critical/dummy classification. As can be seen, APR can lead to an accuracy rate equal to 34.7% (also with $N_{mix} = 4$) which is 34% lower than MFCC based recognition. Results of experiments with supervised critical/dummy classification are showed in Table 6. According to Table 6, the highest accuracy is also achieved with $N_{mix} = 4$ and is equal to 45.7%. This accuracy is substantially higher than the one obtained with the unsupervised procedure, but it is still lower than the accuracy achieved with the MFCC based recognition scheme. It is worth highlighting that the AFs considered here are context independent. Note that the EMA data also provide only partial (i.e., oral cavity) articulatory vocal tract movement information.

Table 4: *Accuracy rate (%) in phoneme recognition by employing the baseline system with MFCC.*

Gaussian mixture number	Accuracy (%)
2	53.09
4	53.11
8	52.15
16	51.12

Table 5: *Accuracy rate (%) in phoneme recognition by employing a classifier based on AFs. The critical and dummy selection is made by means of the automatic stroke classification described in section 2.*

Gaussian mixture number	Accuracy (%)
2	32.89
4	34.76
8	30.03
16	30.47

Table 6: *Accuracy rate (%) in phoneme recognition by employing a classifier based on AFs. The critical and dummy selection is made according to the dictionary presented in [8].*

Gaussian mixture number	Accuracy (%)
2	42.61
4	45.71
8	42.61
16	36.27

Results by combining AS and MFCC features are presented in Table 7. As can be seen, the highest recognition accuracy is reached when $N_{mix} = 2$ and is equal to 61%. The improvement in recognition accuracy with the combined features is as high as 15% and 35% when compared with the MFCC and the AS based systems, respectively. Figure 4 shows the accuracy as a function of the consonants that appeared in the test database where the tongue tip sensor is critical. Consonant 't' shows the highest improvement (75% with respect to the MFCC based phoneme recognition). However, recognition of consonants 'th' and 'sh' seem to be degraded when articulatory information is added.

Experiments by employing MCS approach are presented in Tables 8-9. As can be seen in Table 8, the product rule approximation provides the highest recognition accuracy (equal to 63%) when N_{mix} = 2. Notice that this accuracy is higher than the one shown in Table 7. It could be due to the fact that the joint conditional *a priori* pdf may require more training data to obtain a reliably estimation.

Table 7: *Accuracy rate (%) in phoneme recognition by employing a classifier based on AFs and MFCC. The critical and dummy selection is made according to the dictionary presented in [8].*

Gaussian mixture number	Accuracy (%)
2	61.37
4	58.39
8	54.04
16	47.08

Table 8: *Accuracy rate (%) in phoneme recognition employing the product rule approximation. The critical and dummy selection is made according to the dictionary presented in [8].*

Gaussian mixture number	Accuracy (%)
2	63.06
4	60.37
8	55.03
16	53.29

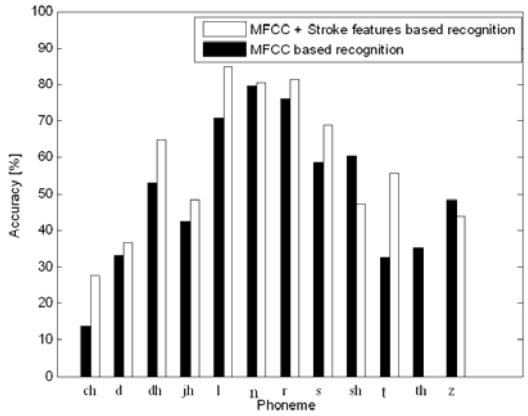

Figure 4: *Accuracy rate (%) in phoneme recognition by employing a classifier based on AFs and MFCC. The critical and dummy selection is made according to the dictionary presented in [8].*

5. Conclusion

This paper presents a promising framework to incorporate articulatory movement information in a classical automatic phoneme recognition scheme. The stroke definition establishes a mechanism to segment any articulatory trajectory. The articulatory strokes are classified as critical or dummy depending on if they are relevant to a specific phoneme. The articulatory stroke classification accuracy can be as high as 88% with the same training and testing speaker. This accuracy decreases by 29% when training and testing data correspond to different speakers. This result is attributed to the fact that the articulatory features depend on the speaker physiological characteristics. This problem could be addressed by means of a speaker normalization method in the articulatory feature domain.

The phonetic recognition accuracy with just articulatory stroke information is still 20% lower than the one achieved with the ordinary MFCC parameters. However, when the phoneme recognition employs both articulatory feature and MFCC vectors, the system accuracy increases by 15% when compared with the baseline system with the MFCC parameters only. This result strongly suggests that the articulatory features provide complementary information with respect to the MFCCs.

Table 9: *Accuracy rate (%) in phoneme recognition employing the mean rule approximation. The critical and dummy selection is made according to the dictionary presented in [8].*

Gaussian mixture number	Accuracy (%)
2	56.15
4	53.79
8	49.81
16	47.45

6. Acknowledgements

This work was funded by a Conicyt-Chile internship grant to visit SAIL at University of Southern California, CA, USA, and Fondecyt No. 1100195 and Fondef No. D05I-10243. The first and last authors thank all the SAIL team for the help provided. The work of the SAIL authors was supported by NSF and NIH.

7. References

[1] Cui X. and Alwan A., (2005) "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR,"IEEE T-SAP,13(6),pp.1161-1172.

[2] Leggetter C., and Woodland P., (1995) "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov Models," Computer Speech and Language. 9(2), pp. 171-185.

[3] Myrvoll T., Siohan O., Lee C.H., and Chou W., (2000) "Structural maximum a posteriori linear Regression for unsupervised speaker adaptation," In ICSLP, vol. 4, pp. 540-543., Beijing, China.

[4] Kirchhoff K., Fink G., and Sagerer G., (2002) "Combining acoustic and articulatory feature information for robust speech recognition," Speech Comm., vol 37, pp.303-319, July.

[5] Tang M., Seneff S., and Zue V., (2003) "Modeling Linguistic Features in Speech Recognition," in European Conference on Speech Comm. and Technology, pp. 2585–2588.

[6] King S., Frankel J., Livescu K., McDermott E., Richmond K., and Wester M., (2007) "Speech production knowledge in automatic speech recognition," Journal of the Acoustical Society of America, vol. 121, pp. 723-742.

[7] Kato T, Lee S. and Narayanan S., (2009) "An analysis of articulatory-acoustic data based on articulatory strokes", ICASSP, pp 4493-4496.

[8] "TADA manual v0.9", http://www.haskins.yale.edu/

[9] Kittler, J., Hatef, M., Duin, R., Matas, J., (1998) "On combining classifiers," IEEE T-PAMI 20, pp. 226-239.

[10] Fumera G. and Roli F., (2005) "A theoretical and experimental analysis of linear combiners for multiple classifier systems," IEEE Trans. Pattern Analysis and Machine Intelligence. 27(6). pp. 942-956.

[11] Wrench A., (2001) "A new source for production modeling in speech technology," Proc. of Workshop on Innovations in Speech Processing.